A Step-by-Step Introduction to Building a Student-at-Risk Prediction Model Using SPSS

http://www.unr.edu/ia/research
http://www.uhwo.hawaii.edu/academics/oie/research-and-presentations/

*Serge Herzog, PhD*
Director, Institutional Analysis
Consultant, CRDA StatLab
University of Nevada, Reno
Reno, NV, serge@unr.edu

*John Stanley, MEd*
Director, Institutional Research
University of Hawaii – West Oahu
Kapolei, HI, jstanley@hawaii.edu

AIR Forum 2017
Washington D.C May 29th – June 2nd
Workshop Objectives

1. Develop a conceptual understanding of how predictive models developed by an IR office can improve institutional effectiveness;

2. Learn how to set up a matriculation system (or census warehouse) data file in SPSS that can be used to develop a predictive statistical model to identify students at risk;

3. Learn how to use historical data to ‘automatically’ develop predictor coefficients to estimate (score) the dropout risk for students in future cohorts; and

4. Learn how to translate the student dropout risk into a relative percentile risk score to assist student support services with ‘actionable’ information.
Two Institutions, One Mission

University of Hawai'i*
West O'ahu

University of Nevada, Reno

PacAIR
Pacific Association for Institutional Research

AIR
Association for Institutional Research

RMAIR
ROCKY MOUNTAIN ASSOCIATION FOR INSTITUTIONAL RESEARCH
Challenges for Institutional Research

- Compliance vs. Self-Improvement
- Developing a culture of evidence
- From reporting to analysis
- Converting results into ‘actionable’ statements
- From ‘data silos’ to integrated warehouse
- Leverage technology, stay abreast of tech
- Follow highest standards, best practices
- Know your customers, mission
- Empower staff, continuous honing of skills
The Institutional Context

- Student success: a strategic imperative
- Performance-based state funding impending
- Dwindling state support for higher education
- Tuition-revenue maximization
- Reputation and marketing
- Effective senior-management support by IR
- K-16 Education Collaborative
  - High school transcript study
  - High school gateway curriculum
  - Reversing the tide of college remediation
The Institutional Context

New Freshmen Enrollment
Examples of Actionable Findings

• Study abroad enhances academic performance

• Impact of classroom facilities/schedule on learning
  – Smaller rooms are preferable
  – After-2pm courses associated with lower performance

• Student financial aid to maximize retention
  – Tuition discounts for middle-income students
  – More academic support for low-income students

• Effect of high school environment on freshmen success
  – http://www.uark.edu/ua/der/EWPA/Research/Achievement/1808.html
Raising Graduation Rates
Comparing 4-year and 6-year-plus Graduates

**Opportunity cost of staying one more year in college = $32,000 in foregone earnings plus annual increase in tuition cost.*

- HS GPA: 3.5 vs 3.2
- ACT: 24.5 vs 22.2
- First-Y GPA: 3.35 vs 2.71
- MathGPA: 3.12 vs 2.4
- Honors Courses: 14% vs 5%
- CoreHum 201 Grade: 3.3 vs 2.6
- Change in Major: 25% vs 55%
- Capstone GPA: 3.5 vs 3.2
- Avg annual remaining need: $2,610 vs $3,270
- Difference in avg semester load: 3 credits
- Final GPA: 3.4 vs 2.9
- Internship: 31% vs 24%

---

Improving the Bottom Line

- Rise in freshmen retention by 4 percentage points due to better at-risk forecasting
  - AY 2010-11 additional net tuition revenues = $215,119 (for 94 NV, 19 WUE, excl OS students) for one cohort in one year, without OS $!
  - Downstream cumulative additional net tuition revenues result in $ millions!

- Incentive for student to speed up graduation
  - Opportunity cost per year in foregone earnings = $32,000 per year (published constant 2010-$)
Relevant Previous Research

Impact of this At-Risk Forecasting Model

- University Retention Rates Hold Steady As States Balance Access with Success. Scripps Howard Foundation Wire, April 15, 2011.


- Consulting services to IR offices at institutions in Arizona, California, Hawaii, and Texas.
At-Risk Forecasting Model

- Identify at-risk freshmen students after initial matriculation for *early* intervention program

- Develop regression model to predict dropout risk of future cohort
  - Determine baseline retention to maximize correct classification
  - Identify statistical outliers to get trimmed dataset
  - Chose model with optimal balance in correct classification

- Dropout risk scoring for new freshmen
  - Transformation of the logit($p$) into probability scores
  - Automated classification and probability score with SPSS
  - Decile grouping of scored students

- Reporting of dropout risk via secure online access
Goal 2: Data file setup

- **Data sources**
  - Matriculation system (Peoplesoft, data warehouse)
  - New student survey (in PS starting fall 2011)

- **Student cohorts**
  - New full-time first-year students (incl. advanced standing)
  - Historical cohorts: fall 2011-15 (training set, N = 4,446)
  - Predicted cohort: fall 2016 (holdout set, N = 986)
  - Excluding ~ 10% of students without entry survey data

- **Data elements (predictors) at start of first semester**
  - Student socio-demographics (personal, parent attributes)
  - Academic preparation (high school GPA, test scores)
  - Financial aid profile (unmet need, aid type received, income)
  - Student motivation (proxy variables)
  - Student social integration (on-campus experiences)
  - Student academic experience (credit load, math/English)
Goal 2: Data file setup

- **Student socio-demographics** (10 predictors)
  - Age19Plus, Male, Hisp, Blk, OS, OSDisc, Non-Local, MotherEd, FatherEd, Pell

- **Academic preparation** (2 predictors)
  - HSPrep (HS Core GPA/Test Score Index), AdvStanding

- **Financial aid profile** (8 predictors)
  - Unmet, Loans, Merit, Inc38827 Inc77464 Inc125776 Inc125776up; FAComplete

- **Student motivation** (2 predictors)
  - EdGoal, FirstChoice

- **Student social integration** (5 predictors)
  - LLC, CampWork, OnCampus, PlanWorkNo, PlanWorkFT

- **Student academic experience** (6 or 7 predictors)
  - Crs13to15, Crs16up, NoEngl, NoMath, DistEd, Undeclared, MidtermGPA (if available)
Data Management Tasks

• Exploratory data analysis
  – Variable selection (bivariate regression on outcome variable)
  – Variable coding (*continuous vs. dummy/binary*)
  – Missing data imputation
  – Derived variable(s)
    • $\text{HSPrep} = (\text{HSGPA} \times 12.5) + (\text{ACTM} \times 0.69) + (\text{ACTE} \times 0.69)$

• Logistic regression model
  – Preliminary model fit (-2LL test/score, pseudo $R^2$, HL sig.)
  – Check for outliers with diagnostic tools (Cook’s, Std Residuals)
  – Check correct classification rate (CCR) for enrollees vs. non-enrollees (i.e. model sensitivity vs. specificity) using baseline probability and Receiver Operating Characteristics (ROC) curve
Data Management Tasks

- Imputation example: HS Preparation index score for cases with missing core GPA or test score
  - Regress core GPA and test score on each other
  - Use regression coefficients to estimate GPA/test score, respectively
  - Run HSPrep index equation for new cases

Table:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.167</td>
<td>.027</td>
<td>79.054</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>ACT_COMP</td>
<td>.060</td>
<td>.001</td>
<td>.419</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>51.618</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: HS_CORE_GPA

© Serge Herzog
Data Management Tasks

- Determine persistence rate of your historical cohorts (fall 2011 through fall 2015): (Set TrainingSpring, TrainingFall = 1)
  - Fall-to-Spring
    - SprRetention
      | Frequency | Percent | Valid Percent | Cumulative Percent |
      |-----------|---------|---------------|--------------------|
      | Valid 0   | 975     | 21.9          | 21.9               |
      | 1         | 3471    | 78.1          | 100.0              |
      | Total     | 4446    | 100.0         | 100.0              |
  - Fall-to-Fall
    - FallRetention
      | Frequency | Percent | Valid Percent | Cumulative Percent |
      |-----------|---------|---------------|--------------------|
      | Valid 0   | 1294    | 29.1          | 29.1               |
      | 1         | 3152    | 70.9          | 100.0              |
      | Total     | 4446    | 100.0         | 100.0              |
Goal 3: Estimate dropout risk

SPSS Menu Tasks

• Select Analyze, Regression, Binary
SPSS Menu Tasks

- Select Analyze, Regression, Binary, Save
Goal 3: Estimate dropout risk

SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Under Options, select HL goodness-of-fit
  - Reset classification cutoff from 0.5 (default) to historical rate
SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Under Selection Variable, select Training variable, click Rule, insert 1
  - Click Paste (inserts syntax in syntax window)

Goal 3: Estimate dropout risk

© Serge Herzog
Goal 3: Estimate dropout risk

**SPSS Menu Tasks**

- Select *Analyze, Regression, Binary*
  - Click Paste (creates syntax in new window)
- Edit syntax as needed to re-specify parameters, re-estimate the dropout risk
- Or use syntax provided in SPSS file

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES SprRetention
  /SELECT=TrainingSpring EQ 1
  /METHOD=ENTER AdvStanding NoMath NoEngl DistEd Undeclared Age19plus Male Hisp Blk OS NonLocal WUE OnCampus CampWork Pell Unmet Loans Merit FirstChoice EdGoalGrad MoEd4yrColl FathEd4yrColl PlanWorkFT PlanWorkNo LLC Crs13to15 Crs16up HSPrep Inc38827 Inc77464 Inc125776 Inc125776up FAComplete
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.781).
```

© Serge Herzog
SPSS Output File

• Correct classification rate (CCR) for historical data is ~65%, for fall 2016 cohort it is ~66%.
• To improve CCR, check and exclude outlier cases

Goal 3: Estimate dropout risk

Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Selected Cases</th>
<th>Predicted</th>
<th>Unselected Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>Percentage Correct</td>
<td>0</td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SprRetention</td>
<td>637</td>
<td>338</td>
<td>65.3</td>
<td>102</td>
</tr>
<tr>
<td>1</td>
<td>1235</td>
<td>2236</td>
<td>64.4</td>
<td>240</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td></td>
<td>64.6</td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .781
b. Selected cases TrainingSpring EQ 1
c. Unselected cases TrainingSpring NE 1
Identify and Exclude Outlier Cases

- Exclude Mahalanobis Distance (optional)
- Examine Cook’s distance (COO) and standardized residuals (ZRE) for training data

Exclude cases with
- Cook’s distance greater than 1, or visual separation
- Standardized residuals greater than |3|

More stringent exclusion rules
- Cook’s distance greater than $\frac{4}{n} = \text{number of cases}$
- Standardized residuals greater than |2|
Excluding Outlier Cases

Goal 3: Estimate dropout risk

ZRE_1 > -3 | Term = ‘F16’
Results from Trimmed Data

- Cut value adjusted to .792 to reflect trimmed training data
- Overall CCR at ~67% both historical and fall 2016 cohorts
- R-square = .21, but HL reached significance (<.05)
- Improve CCR by including Mid-Term Grades

Goal 3: Estimate dropout risk

Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Selected Cases</th>
<th>Predicted</th>
<th>Unselected Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SprRetention</td>
<td>Percentage Correct</td>
<td>SprRetention</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>621</td>
<td>291</td>
<td>68.1</td>
</tr>
<tr>
<td></td>
<td>1164</td>
<td>2307</td>
<td>66.5</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>66.8</td>
</tr>
</tbody>
</table>

a. The cut value is .792
b. Selected cases TrainingSpring EQ 1
c. Unselected cases TrainingSpring NE 1
Results with Mid-Term Grades

- Include ‘mid term’ variable in syntax window
- Select all cases, no outlier exclusions: Cut value at 0.781
- Overall CCR at 82% for fall 2016 cohort
- R-square = .44, but HL reached significance (<.05)
- BUT, mediocre CCR for fall 2016 dropout students (58.6%)

Goal 3: Estimate dropout risk

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Selected Cases</th>
<th>Predicted</th>
<th>Unselected Cases</th>
<th>Percentage Correct</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>706</td>
<td>269</td>
<td>109</td>
<td>77</td>
<td>58.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>618</td>
<td>2853</td>
<td>99</td>
<td>687</td>
<td>87.4</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>80.0</td>
<td></td>
<td></td>
<td></td>
<td>81.9</td>
</tr>
</tbody>
</table>

a. The cut value is .781
b. Selected cases TrainingSpring EQ 1
c. Unselected cases TrainingSpring NE 1
d. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
Results with Mid-Term Grades

- Select all cases, no outlier exclusions: Cut value at 0.781
- **Change classification cutoff value to 0.87**
- Overall CCR down (72.4%), but more balanced CCR
- Nearly 70% CCR for dropout cases in predicted (fall ‘16) cohort

---

Goal 3: Estimate dropout risk

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>SprRetention</th>
<th>Selected Cases</th>
<th>Percentage Correct</th>
<th>Predicted</th>
<th>SprRetention</th>
<th>Unselected Cases</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>SprRetention</td>
<td>795</td>
<td>180</td>
<td>81.5</td>
<td>128</td>
<td>58</td>
<td>68.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2355</td>
<td>67.8</td>
<td>210</td>
<td>576</td>
<td>73.3</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>1116</td>
<td>2355</td>
<td>67.8</td>
<td>70.9</td>
<td></td>
<td>72.4</td>
<td></td>
</tr>
</tbody>
</table>

- The cut value is .870
- Selected cases TrainingSpring EQ 1
- Unselected cases TrainingSpring NE 1
- Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

© Serge Herzog
Pondering Results

• Outlier removal improves prediction accuracy, but exclusion of too many cases may bias results
• Midterm prediction, including midterm grades, yields higher prediction accuracy without exclusion of outlier cases
• Thus, prediction accuracy is a balancing act between waiting for more pertinent data (e.g. midterm grades) and excluding outlier cases for better model fit but possibility of biasing results
• When excluding outlier cases, examine how many are removed (keep number of excluded outliers below 5% of total cases; check coefficient of determination R-square, Hosmer-Lemeshow alpha level preferably > 0.05)
Determine Balanced CCR: ROC Charts

Goal 3: Estimate dropout risk

Test Variable:
- Predicted probability [FRE_3]

State Variable:
- SprRetention
  Value of State Variable: 1

Display:
- ROC Curve
- With diagonal reference line
- Standard error and confidence interval
- Coordinate points of the ROC Curve

© Serge Herzog
Determine Balanced CCR: ROC Charts

- Simultaneous measure of sensitivity (true positive) and specificity (true negative) for all possible cutoff values
- Calculate area under the ROC curve (exercise)
- Area under the ROC: .901 (all case data)
- Suggested cutoff point to maximize overall CCR is around 0.901. (associated CCR for dropout = 73.1%)
Assess Prediction Accuracy

- Compare results from full-data model with results from trimmed-data model
- Determine the best cut value (classification) based on re-adjusted baseline probability versus ROC-curve derived probability level
- Evaluate relative cost of (in-)accurate prediction of retained students (sensitivity) versus dropout students (specificity)
- Usually, err in favor of accurate identification of students at risk of dropping out, without sacrificing too much accuracy for retained
Translate Dropout Risk

• Copy retention probability for fall 2016 cohort to new file (to eliminate all other cases)
• Group into deciles using binning function:
  – Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
• Note bottom high-risk deciles with far lower retention probability (run decile average)
• Identify cusp of probability border between predicted dropouts and persisters and corresponding decile groups
• Identify priority decile groups near the cusp for student assistance
## Sample Data for Advisors

<table>
<thead>
<tr>
<th>R Number</th>
<th>Last Name</th>
<th>First Name</th>
<th>Email Addr</th>
<th>Age</th>
<th>College</th>
<th>Dept</th>
<th>Major</th>
<th>Dropout Risk Decile</th>
<th>Relative Spring Retention %tile</th>
</tr>
</thead>
<tbody>
<tr>
<td>18LBA</td>
<td>ART</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>14.92</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>28.52</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>36.80</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>39.18</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>46.87</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>66.48</td>
</tr>
<tr>
<td>19LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>92.42</td>
</tr>
<tr>
<td>18LBA</td>
<td>ANTH</td>
<td></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>95.57</td>
</tr>
</tbody>
</table>
### Sample Data for Advisors

<table>
<thead>
<tr>
<th>Gender</th>
<th>Ethnicity</th>
<th>Credits</th>
<th>Resident</th>
<th>State/Cnty</th>
<th>HS GPA</th>
<th>ACTE</th>
<th>ACTM</th>
<th>Has Pell$ (1=yes)</th>
<th>Has Loan$ (1=yes)</th>
<th>Clark Cnty Resi (1=yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>AS</td>
<td>12 NV</td>
<td>NWA</td>
<td></td>
<td>3.10</td>
<td>24</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>15 NV</td>
<td>NCL</td>
<td></td>
<td>3.23</td>
<td>21</td>
<td>18</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>16 WU</td>
<td>CA</td>
<td></td>
<td>3.19</td>
<td>23</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>17 WU</td>
<td>OR</td>
<td></td>
<td>3.23</td>
<td>24</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>16 NV</td>
<td>NWA</td>
<td></td>
<td>3.18</td>
<td>17</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>WH</td>
<td>15 NV</td>
<td>NDO</td>
<td></td>
<td>3.47</td>
<td>30</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>WH</td>
<td>15 NV</td>
<td>NWA</td>
<td></td>
<td>3.65</td>
<td>26</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>AS</td>
<td>16 NV</td>
<td>NCL</td>
<td></td>
<td>3.90</td>
<td>30</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Unbalanced Data

- Proportion of dropouts is usually much smaller than proportion of retained students.
- Number of cases in rare event (dropout) should be sufficient to yield minimum 10:1 ratio with number of predictors (preferably 30:1 ratio).
- Check standard errors in coefficient results table ("Variables in the Equation") for inflated values.
- Check variance inflation factor (VIF) in collinearity diagnostics (must run linear regression) to determine which predictor(s) to remove if ratio well below 10:1 or run Exact Logistic Regression (see example at http://www.ats.ucla.edu/stat/stata/dae/exlogit.htm).
Exercise

• Estimate fall-to-fall dropout risk for 2016 cohort, using 2011 through 2015 cohorts
Exercise

- Estimate fall-to-fall dropout risk for 2016 cohort using 2011 through 2015 cohorts
- Set cutoff value = 0.709. All cases included.
- Check/exclude outliers, re-run model

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>FallRetention</th>
<th>Percentage Correct</th>
<th>Unselected Cases</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>0</td>
<td>882</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>643</td>
<td>2509</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The cut value is .709
- Selected cases TrainingFall EQ 1
- There are no unselected cases. Therefore, no unselected cases are classified.
- Unselected cases TrainingFall NE 1
- Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
Exercise

• Excluding Z-residuals (> +/-3), 95 cases (2.1%)
• CCR improved to 78.6% from 76.3%
• R-square 0.47 (cut value adjusted to .723)

Goal 5: Estimate fall-fall dropout risk

Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Selected Cases</th>
<th>Predicted</th>
<th>Unselected Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FallRetention 0</td>
<td>Percentage Correct</td>
<td>FallRetention 0</td>
</tr>
<tr>
<td>Step 1</td>
<td>0</td>
<td>876</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>599</td>
<td>81.0</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>78.6</td>
<td>1075</td>
<td>.</td>
</tr>
</tbody>
</table>

a. The cut value is .723
b. Selected cases TrainingFall EQ 1
c. There are no unselected cases. Therefore, no unselected cases are classified.
d. Unselected cases TrainingFall NE 1
e. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
Translate Dropout Risk

- Copy retention probability for fall 2016 cohort to new file (to eliminate all other cases)
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
- Note bottom high-risk deciles with far lower retention probability (run decile average)
- Identify cusp of probability border between predicted dropouts and persisters and corresponding decile groups
- Identify priority decile groups near the cusp for student assistance
- Send student record file with predicted probability, predicted outcome, decile group to student assistance/advising personnel
Community College Data Set Details

Mimic* dataset based on data from:

- ~ 4,300 student enrollment
- Open access
- Large % of under-represented, low income, and first generation students
- 60% male
- Average age is 26 years old
- 66% part-time enrollment
- Over half of academic programs are vocational/career technical
- 18% grad rate (150%)
- 72% fall-to-spring retention first-time freshmen; 50% fall-to-fall retention

*The CC Dataset used in this class has been de-identified, randomized, and altered for instructional and sharing purposes. These “mimic” data do not match actual institutional data, trends, or outcomes.
Community College Data Set Details

• Data Sources
  – Matriculation system (Banner, data warehouse)

• Student cohorts
  – New first-year students (part-time and full-time)
  – Historical cohorts: fall 2013-15 (training set, N=2,243)
  – Predicted cohort: fall 2016 (holdout set, N=626)
  – Newest cohort: fall 2017 (holdout set #2, N=702)

• Data elements (predictors) at start of first semester
  – Student socio-geo-demographics (age, gender, ethnicity, geographic proximity to campus, residency, military)
  – Academic preparation (Compass test scores, high school attended, remediation/ developmental courses needed)
  – Financial aid profile (unmet need, pell)
  – Student motivation proxies (degree audit logins, educational goals survey responses)
  – Student academic experience (credit load, math/English enrollment, major type)
Goal 2: Data File Setup

35 predictor variables in the data set

- Student socio-demographics (12 predictors)
  - AGE, AGE19PLUS, FEMALE, URM, URMINCFILIPINO, WHITE, ISLANDWEST, ISLANDURBAN, ISLANDRURAL, OUTOFSTATE, MILITARY, LOWPERFORMHIGHSCHOOL
- Academic preparation (9 predictors)
  - COMPASS READING, COMPASS WRITING, COMPASSANYMATHHIGHEST, REMEDIAL/DEVELOPMENTAL/COLLEGELEVEL (Math/English) FLAGS,
- Financial aid profile (2 predictors)
  - PERCENTUNMETNEED, PELL
- Student motivation (4 predictors)
  - EDGOAL1, EDGOAL2, STARUSAGE, STARUSAGEAVERAGEFLAG,
- Student academic experience (8 predictors)
  - CREDITSATTEMPTED, CREDITSLESS9, FULLTIME, DISTANCEEDENROLL, ECED MAJOR, APPLIEDTRADESMAJOR, ANYMATHENROLL, ANYENGLISHENROLL
Step 1: Filter out the 2015 data

Select Data, Select Cases, *If condition…*

COHORTYEAR ~ = 2017
Goal 3: Estimate dropout risk

CC Data: SPSS Menu Tasks

- Select Analyze, Regression, Binary
  - Use same menu options learned in the UNR example.
  - Click Paste (creates syntax in new window).
- From here on, we will edit syntax as needed to re-specify parameters, re-estimate the dropout risk

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE
  LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING
  COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```
Goal 3: Estimate dropout risk

SPSS Output File

- R-square = .255 ; HL test sig. = .103
- Null model correct classification rate (CCR) for spring dropout is nil in both training and holdout data (0.0%)

Here, we calculated the baseline fall-to-spring retention rate
SPSS Menu Tasks

- Select *Analyze, Regression, Binary*
  - Click Paste (creates syntax in a new window)
- Edit cut value in syntax to reflect baseline probability of spring retention (i.e. 72.1%)

```spss
DATASET ACTIVATE DataSet2.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTED DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG /SAVE=PRED PGROUP COOK ZRESID /PRINT=GOODFIT /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) **CUT(0.721)**.
```
SPSS Output File

- R-square = .255; HL test sig. = .103
- CCR for spring dropout at 70% for training and 80% for holdout cohorts
- Good correct classification rate of dropout students
  - Check for outliers to seek further improvement

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage Correct</th>
<th>Selected Cases</th>
<th></th>
<th></th>
<th>Unselected Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RETENTIONSPRING</td>
<td></td>
<td>RETENTIONSPRING</td>
<td>0</td>
<td>1</td>
<td>RETENTIONSPRING</td>
</tr>
<tr>
<td>Step 1</td>
<td>RETENTIONSPRING</td>
<td></td>
<td>440</td>
<td>186</td>
<td>70.3</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>RETENTIONSPRING</td>
<td></td>
<td>490</td>
<td>1127</td>
<td>69.7</td>
<td>164</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
a. The cut value is .721
b. Selected cases TRAINING_VARIABLE EQ 1
c. Unselected cases TRAINING_VARIABLE NE 1
Identify Outlier Cases

- Examine Cook’s distance (COO\_) and standardized residuals (ZRE\_)
- Exclude cases with
  - Cook’s distance greater than 1, or visual separation
  - Standardized residuals greater \(|3|\)
- More stringent exclusion rules
  - Cook’s distance greater than \(4/n=\text{number of cases}\)
  - Standardized residuals greater \(|2|\)
Goal 3: Estimate dropout risk

Identify Outlier Cases

Cook’s Values of 0.1 or higher merit outlier exclusion from “eyeballing” the scatterplot.
SPSS Menu Tasks

- Exclude outliers via ‘select cases if’ function
- Use ‘filter_Trim (already included)

COHORTYEAR ~= 2017 & (COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > -3)
SPSS Syntax Version of Filter Tasks (fyi)

DATASET ACTIVATE DataSet1.
USE ALL.
COMPUTE filter_$=(COHORTYEAR ~= 2017 & COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > - 3).
VARIABLE LABELS filter_$ 'COHORTYEAR ~= 2017 & (COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > - 3) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
Goal 3: Estimate dropout risk

CC Data: SPSS Menu Tasks

- Run regression syntax again with the 0.721 baseline retention rate

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPPRING
  /SELECT=TRAININGVARIABLE EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE
  LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG
  /SAVE=PRED PGROUP COOK ZRESID
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.721).
```
Goal 3: Estimate dropout risk

Calculate new baseline from trimmed data

- New baseline retention rate is .724 based on trimmed training data

<table>
<thead>
<tr>
<th>Classification Table$^{a,b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Step 0 RETENTIONSPRING</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Overall Percentage</td>
</tr>
</tbody>
</table>

a. Constant is included in the model.
b. The cut value is .721
c. Selected cases TRAININGVARIABLE EQ 1
d. Unselected cases TRAININGVARIABLE NE 1
CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the new baseline retention rate = 0.724

```
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
   /SELECT=TRAININGVARIABLE EQ 1
   /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGEAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATH HIGHEST REMEDIALMATH REMEDIALENG
   /SAVE=PRED PGROUP COOK ZRESID
   /PRINT=GOODFIT
   /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.724).
```
Updated Results from Trimmed Data

- Cut value adjusted to .724 to reflect trimmed training data
- Dropout CCR at 72% for training, 82% for holdout data
- Overall CCR at ~70% for both training and holdout data
- R-square = .295, but HL reached significance (<.05)

Goal 3: Estimate dropout risk

81.5% accuracy in identifying dropped students
Results from Trimmed Data

- Some false positives in Decile 1 for predicting retainers, but overall results suggest stability.

### Contingency Table for Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th></th>
<th>RETENTIONSPRING = 0</th>
<th>RETENTIONSPRING = 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Expected</td>
<td>Observed</td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>144</td>
<td>159.891</td>
<td>79</td>
</tr>
<tr>
<td>2</td>
<td>123</td>
<td>112.615</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>87.850</td>
<td>133</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>70.315</td>
<td>146</td>
</tr>
<tr>
<td>5</td>
<td>53</td>
<td>56.206</td>
<td>170</td>
</tr>
<tr>
<td>6</td>
<td>52</td>
<td>45.928</td>
<td>171</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>36.500</td>
<td>185</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>27.224</td>
<td>189</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>13.452</td>
<td>220</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>4.021</td>
<td>218</td>
</tr>
</tbody>
</table>
Goal 3: Estimate dropout risk

Results from Trimmed Data

- Parameter estimates results. 9 variables significant at .05 level.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CREDITSATTEMPTEDFALL</td>
<td>.152</td>
<td>.016</td>
<td>87.617</td>
<td>1</td>
<td>.000</td>
<td>1.104</td>
</tr>
<tr>
<td>DISTANCEEDENROLLMENT</td>
<td>-.349</td>
<td>.190</td>
<td>3.402</td>
<td>1</td>
<td>.065</td>
<td>.706</td>
</tr>
<tr>
<td>URM</td>
<td>-.643</td>
<td>.113</td>
<td>32.390</td>
<td>1</td>
<td>.000</td>
<td>.526</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-.002</td>
<td>.119</td>
<td>.209</td>
<td>1</td>
<td>.604</td>
<td>.940</td>
</tr>
<tr>
<td>ISLANDRURAL</td>
<td>-.590</td>
<td>.201</td>
<td>3.629</td>
<td>1</td>
<td>.003</td>
<td>.554</td>
</tr>
<tr>
<td>OUTOFSTATE</td>
<td>-.598</td>
<td>.268</td>
<td>4.998</td>
<td>1</td>
<td>.025</td>
<td>.550</td>
</tr>
<tr>
<td>LOWPERFORMHIGHSCHOOL</td>
<td>-.472</td>
<td>.143</td>
<td>10.879</td>
<td>1</td>
<td>.001</td>
<td>.624</td>
</tr>
<tr>
<td>ECEDMAJOR</td>
<td>.022</td>
<td>.272</td>
<td>.007</td>
<td>1</td>
<td>.935</td>
<td>1.023</td>
</tr>
<tr>
<td>AGE19PLUS</td>
<td>-.287</td>
<td>.118</td>
<td>5.012</td>
<td>1</td>
<td>.015</td>
<td>.751</td>
</tr>
<tr>
<td>EDGOAL1</td>
<td>1.655</td>
<td>.265</td>
<td>39.002</td>
<td>1</td>
<td>.000</td>
<td>5.235</td>
</tr>
<tr>
<td>PELL</td>
<td>2.978</td>
<td>.271</td>
<td>120.323</td>
<td>1</td>
<td>.000</td>
<td>19.639</td>
</tr>
<tr>
<td>PERCENTUNMETNEED</td>
<td>-.428</td>
<td>.428</td>
<td>105.951</td>
<td>1</td>
<td>.000</td>
<td>.012</td>
</tr>
<tr>
<td>STARUSAGE</td>
<td>.018</td>
<td>.022</td>
<td>.702</td>
<td>1</td>
<td>.402</td>
<td>1.019</td>
</tr>
<tr>
<td>COMPASSREADING</td>
<td>-.002</td>
<td>.002</td>
<td>.420</td>
<td>1</td>
<td>.517</td>
<td>.998</td>
</tr>
<tr>
<td>COMPASSANYMATHHIGHEST</td>
<td>.005</td>
<td>.003</td>
<td>1.818</td>
<td>1</td>
<td>.178</td>
<td>1.005</td>
</tr>
<tr>
<td>REMEDIALMATH</td>
<td>-.164</td>
<td>.126</td>
<td>1.658</td>
<td>1</td>
<td>.196</td>
<td>.849</td>
</tr>
<tr>
<td>REMEDIALENG</td>
<td>-.216</td>
<td>.136</td>
<td>2.507</td>
<td>1</td>
<td>.113</td>
<td>.808</td>
</tr>
<tr>
<td>Constant</td>
<td>-.276</td>
<td>.284</td>
<td>.945</td>
<td>1</td>
<td>.331</td>
<td>.759</td>
</tr>
</tbody>
</table>

<sup>a</sup> Variable(s) entered on step 1: CREDITSATTEMPTEDFALL, DISTANCEEDENROLLMENT, URM, FEMALE, ISLANDRURAL, OUTOFSTATE, LOWPERFORMHIGHSCHOOL, ECEDMAJOR, AGE19PLUS, EDGOAL1, PELL, PERCENTUNMETNEED, STARUSAGE, COMPASSREADING, COMPASSANYMATHHIGHEST, REMEDIALMATH, REMEDIALENG.
Goal 3: Estimate dropout risk

Final Step: “Go Live” and score the incoming cohort

• Update filter in menu: Select Data, Select Cases, If condition...

In the syntax, change “~=” to “=” for “COHORTYEAR...”

COHORTYEAR = 2017 | (COO_3 < .1 & ZRE_3 < 3 & ZRE_3 > - 3)
CC Data: SPSS Menu Tasks

- Re-run regression syntax AGAIN with the last baseline retention rate = 0.724
- Change “TRAININGVARIABLE2 EQ 1” to score the 2017 cohort.

```spss
DATASET ACTIVATE DataSet1.
LOGISTIC REGRESSION VARIABLES RETENTIONSPRING
  /SELECT=TRAININGVARIABLE2 EQ 1
  /METHOD=ENTER CREDITSATTEMPTEDFALL DISTANCEEDENROLLMENT URM FEMALE ISLANDRURAL OUTOFSTATE LOWPERFORMHIGHSCHOOL ECEDMAJOR AGE19PLUS EDGOAL1 PELL PERCENTUNMETNEED STARUSAGE COMPASSREADING COMPASSANYMATHHIGHEST REMEDIALMATH REMEDIALENG /SAVE=PRED PGROUP COOK ZRESID /PRINT=GOODFIT /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.724).
```
As you see, we generated predicted probability scores (PRE_2) and Group Membership (PGR_2) for the 2015 cohort.
Last Step: Translate Dropout Risk into deciles for easier interpretation by academic support office

- Convert retention probability to dropout risk deciles (1 = highest, 10 = lowest)
- Filter Data for “2017” cohort only.
- Group into deciles using binning function:
  - Transform, Visual Binning, Make 9 cutpoints, Label ‘Deciles’, check ‘reverse scale’
- Note bottom high-risk deciles with far lower retention probability (run decile average)
Goal 4: Assist Student Support

Last Step: Filter 2017 cohort and create new dataset

COHORTYEAR = 2015

Copy selected cases to a new dataset; give it a name.
Goal 4: Assist Student Support

Last Step: Group into deciles using binning function:

- Transform, Visual Binning, Make 9 cutpoints on "PRE_2", Label 'Deciles', check 'reverse scale'
Now your new 2017 dataset has 10 deciles with an even distribution of low-to-high risk scores. Decile 10 is the highest risk.
Goal 4: Assist Student Support

Now that your data is ready, create a spreadsheet for delivery to your advisors/success coaches. Here is an example:

<table>
<thead>
<tr>
<th>ID</th>
<th>LAST NAME</th>
<th>FIRST NAME</th>
<th>EMAIL</th>
<th>CURRENT CREDITS</th>
<th>RESIDENT</th>
<th>AP/CLEP</th>
<th>HS GPA</th>
<th>WORK ON CAMP</th>
<th>1st YR EXP CLASS</th>
<th>% FIN NEED MET</th>
<th>STAR LOGINS</th>
<th>ADVISOR PREVIOUS CONTACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>HI</td>
<td>6</td>
<td>3.80</td>
<td>Y</td>
<td>Y</td>
<td>77%</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>002</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>HI</td>
<td>0</td>
<td>3.33</td>
<td>N</td>
<td>Y</td>
<td>63%</td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>003</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>CA</td>
<td>6</td>
<td>3.00</td>
<td>N</td>
<td>N</td>
<td>45%</td>
<td>0</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>AGE</th>
<th>GENDER</th>
<th>ETHNICITY</th>
<th>COLLEGE</th>
<th>MAJOR</th>
<th>DEGREE</th>
<th>Ed Goal Specified</th>
<th>Relative Risk Value</th>
<th>Risk Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>18</td>
<td>F</td>
<td>CH</td>
<td>CA&amp;H</td>
<td>ART</td>
<td>BA</td>
<td>Yes</td>
<td>14.92</td>
<td>LOW</td>
</tr>
<tr>
<td>002</td>
<td>18</td>
<td>F</td>
<td>HW</td>
<td>CSS</td>
<td>SOC</td>
<td>BA</td>
<td>Yes</td>
<td>36.88</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>003</td>
<td>18</td>
<td>M</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>UNDEC</td>
<td>No</td>
<td>89.18</td>
<td>HIGH</td>
</tr>
</tbody>
</table>
Progress on Implementation at Honolulu Community College (2014)

- Delivering student dropout risk scores to HCC’s Academic Success Center (via an Excel file).
- Training staff members on using the data.
- Academic advisors moving towards a proactive, targeted approach.
Summary

• Predicting students at-risk
  – Keep prediction model parsimonious
  – Keep prediction data for student advising intuitive and simple (actionable)
  – Triangulate prediction data with multiple sources of information
  – Use prediction data as component part of student dropout-risk assessment
  – Follow ‘best practices’ in IR and keep abreast of changes in analytical and data reporting tools

• Using prediction data for student advising
  – Embrace the use of available data
  – Ensure users conceptually understand what’s behind the data
  – Use data as a complementary piece of information when advising students
  – Timing can be critical in terms of student intervention as well as maximizing advising resources

• Stay abreast of new research on predictive analytics:

******************